Pakiet fitdistrplus

Mateusz Kobyłka

Wydział Matematyki i Nauk Informacyjnych
Politechnika Warszawska

Programowanie i analiza danych w R
dla zaawansowanych
Warszawa, 8 maja 2018r.

Plan prezentacji

Wprowadzenie

Pakiet **fitdistrplus** umożliwia dopasowanie rozkładu do danych jednowymiarowym różnymi metodami. Są to:

- metoda największej wiarogodności (MLE),
- metoda momentów (MME),
- metoda kwantyli (MQE),
- metoda najlepszego dopasowania (MGE).

Ponadto, pakiet umożliwia dopasowanie parametrycznego rozkładu i estymację parametrów dla danych cenzurowanych (prawostronnie, lewostronnie bądź przedziałowo).

Wstępna analiza

Pakiet zawiera funkcje, które umożliwiają wstępną analizę danych i wytypowanie rodziny rozkładów, z której mogą pochodzić obserwacje. Są to:

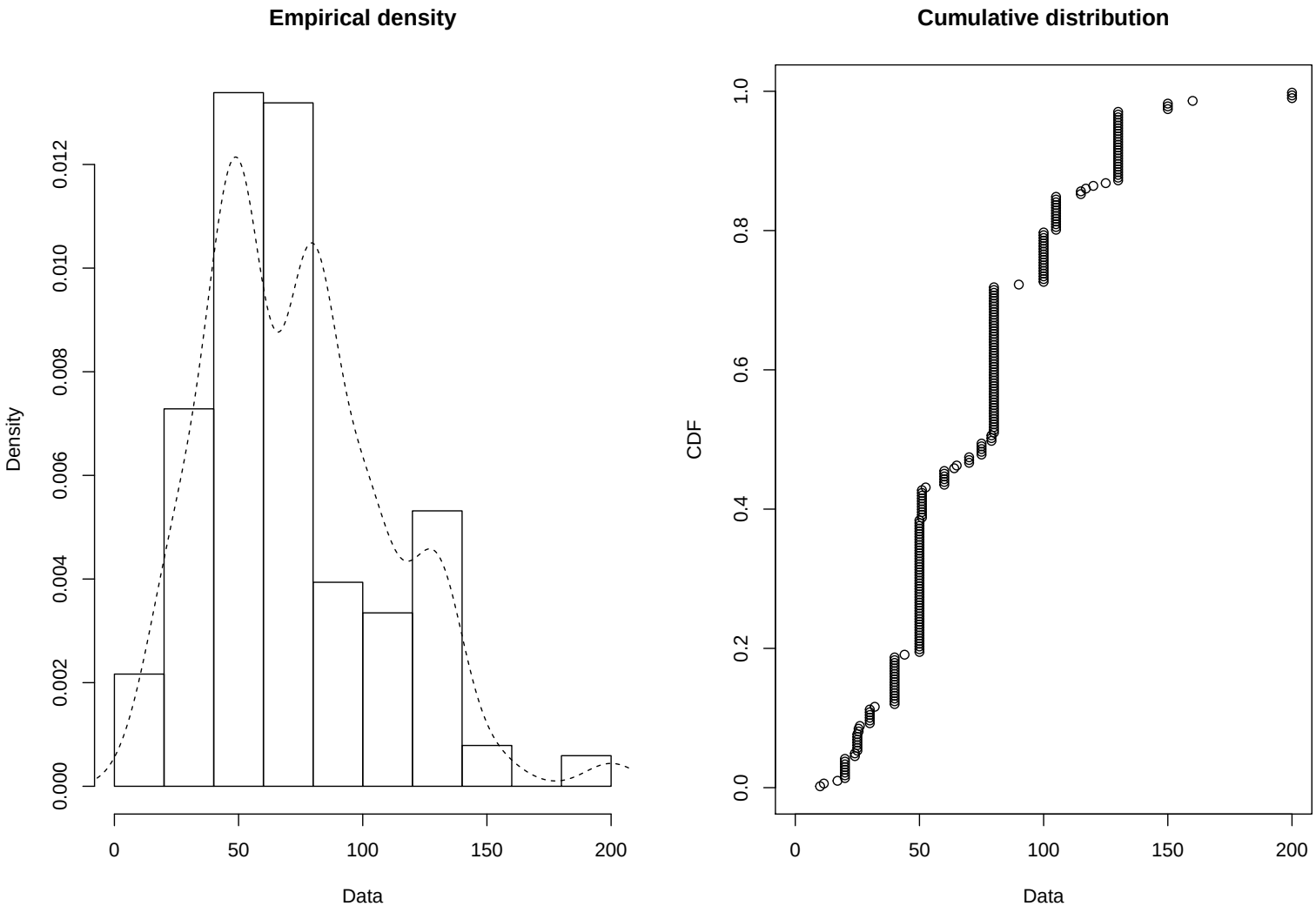
- **plotdist** - rysująca histogram oraz dystrybuantę empiryczną,
- **descdist** - zwracająca podstawowe statystyki oraz graf Cullena-Freya.

Wstępna analiza

Rozważmy zbiór danych **groundbeef** zawarty w pakiecie **fitdstrplus**. Dane zawierają liczbę gramów mielonego mięsa wołowego spożywanego przez dzieci poniżej 5 roku życia we Francji. Na podanym zbiorze danych wywołujemy funkcję **plotdist** oraz **descdist**.

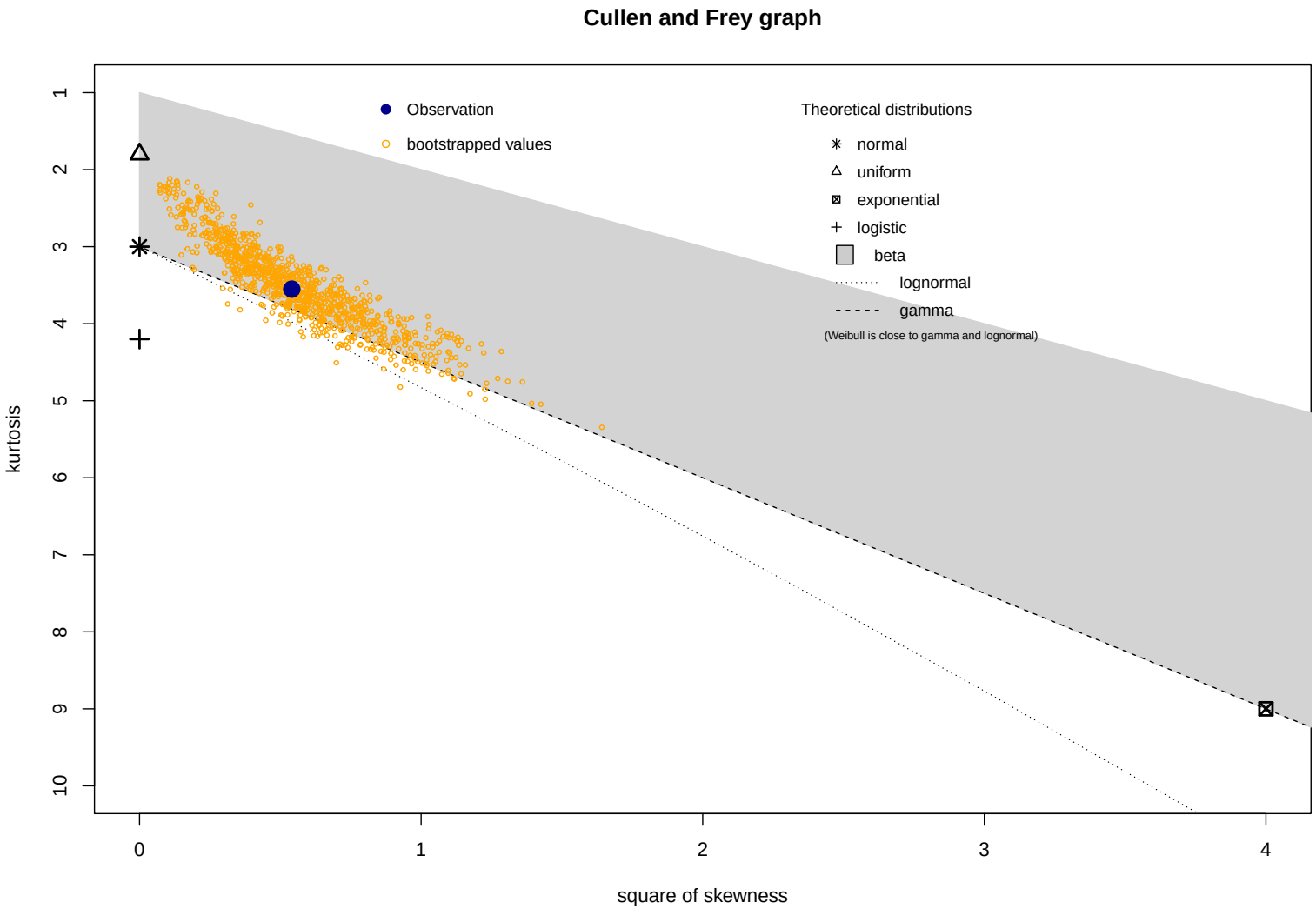
```
library("fitdistrplus")
data("groundbeef", package = "fitdistrplus")
plotdist(groundbeef$serving, histo = TRUE, demp = TRUE)
descdist(groundbeef$serving, boot = 1000)
```

Wstępna analiza



Rysunek 1: Wynik wywołania funkcji **plotdist**

Wstępna analiza



Rysunek 2: Wynik wywołania funkcji **descdist**

Wstępna analiza

Oprócz rysunku, funkcja **descdist** zwraca również wydruk ze statystykami dla podanych danych.

```
summary statistics
------
min:  10   max:  200
median:  79
mean:  73.65
estimated sd:  35.88
estimated skewness:  0.7353
estimated kurtosis:  3.551
```

Dopasowanie rozkładu

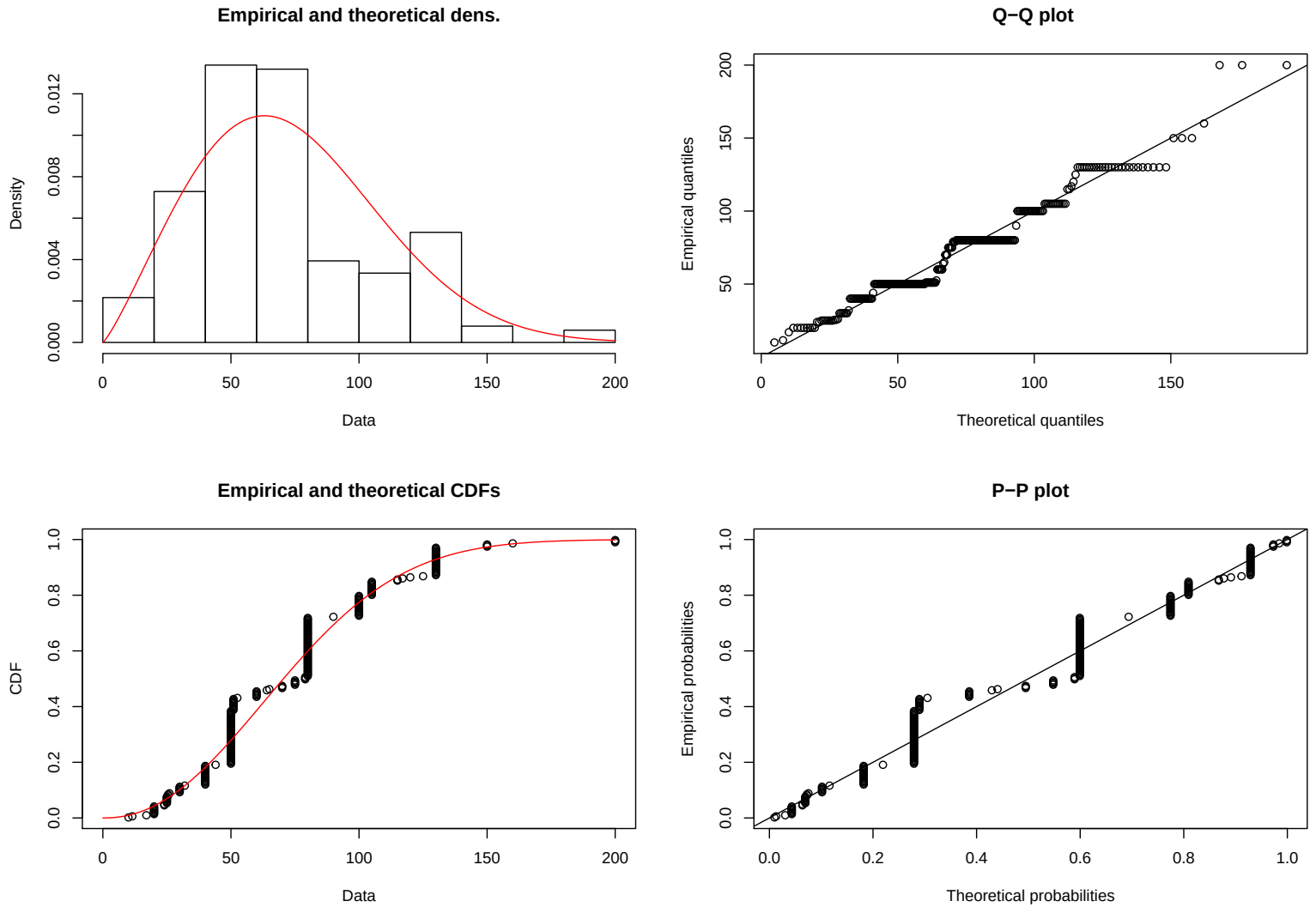
Po wstępnej analizie do danych dopasowany zostanie rozkład Weibulla. Oszacowanie parametrów rozkładu umożliwia funkcja **fitdist**. Funkcja oprócz tego zwraca:

- błędy standardowe wyestymowanych parametrów,
- wartość funkcji logwiarogodności,
- wartości AIC oraz BIC,
- macierz korelacji estymowanych parametrów.

Dopasowanie rozkładu

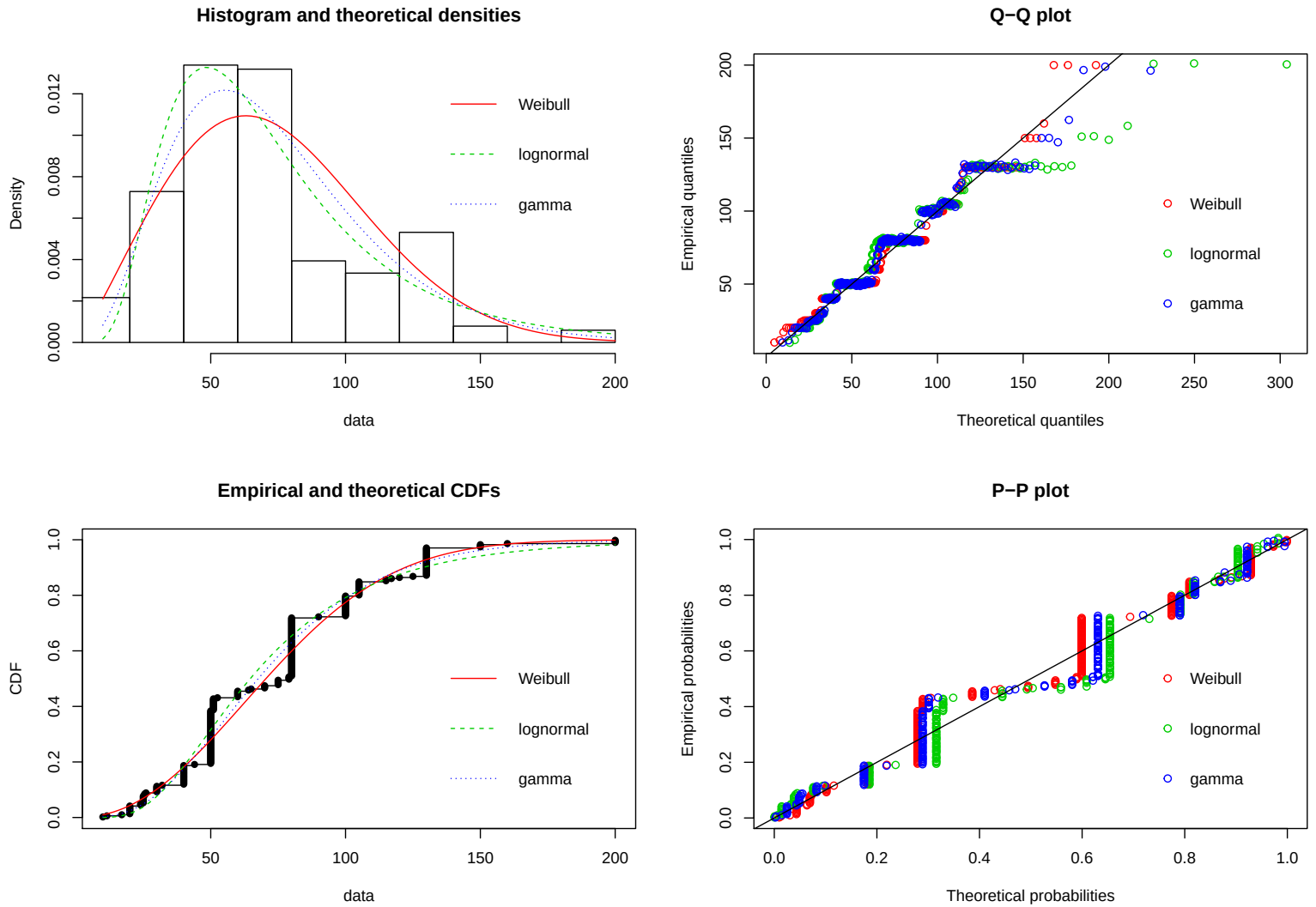
```
fw <- fitdist(groundbeef$serving, "weibull")
summary(fw)

Fitting of the distribution ' weibull ' by maximum likelihood
Parameters :
      estimate Std. Error
shape    2.186     0.1046
scale   83.348     2.5269
Loglikelihood:  -1255   AIC:  2514   BIC:  2522
Correlation matrix:
       shape  scale
shape 1.0000 0.3218
scale 0.3218 1.0000
```

Dopasowanie rozkładu



Rysunek 3: Wynik wywołania **plot(fw)**

Dopasowanie rozkładu



Rysunek 4: Dopasowanie innych rozkładów.

Dane cenzurowane

Czasami zdarza się, że nie posiadamy pełnej informacji o obserwacji, a wiemy tylko, że jest np. mniejsza lub większa od pewnej wartości. Takie dane nazywamy danymi **cenzurowanymi**.

Dane te są dość powszechne np. w analizie przeżycia czy badaniach klinicznych. Dopasowanie rozkładu do nich jest trudniejsze, ale pakiet **fitdistrplus** zawiera funkcje przygotowane specjalnie do tego zadania funkcje.

Dane cenzurowane

Wyróżniamy trzy typy cenzurowań:

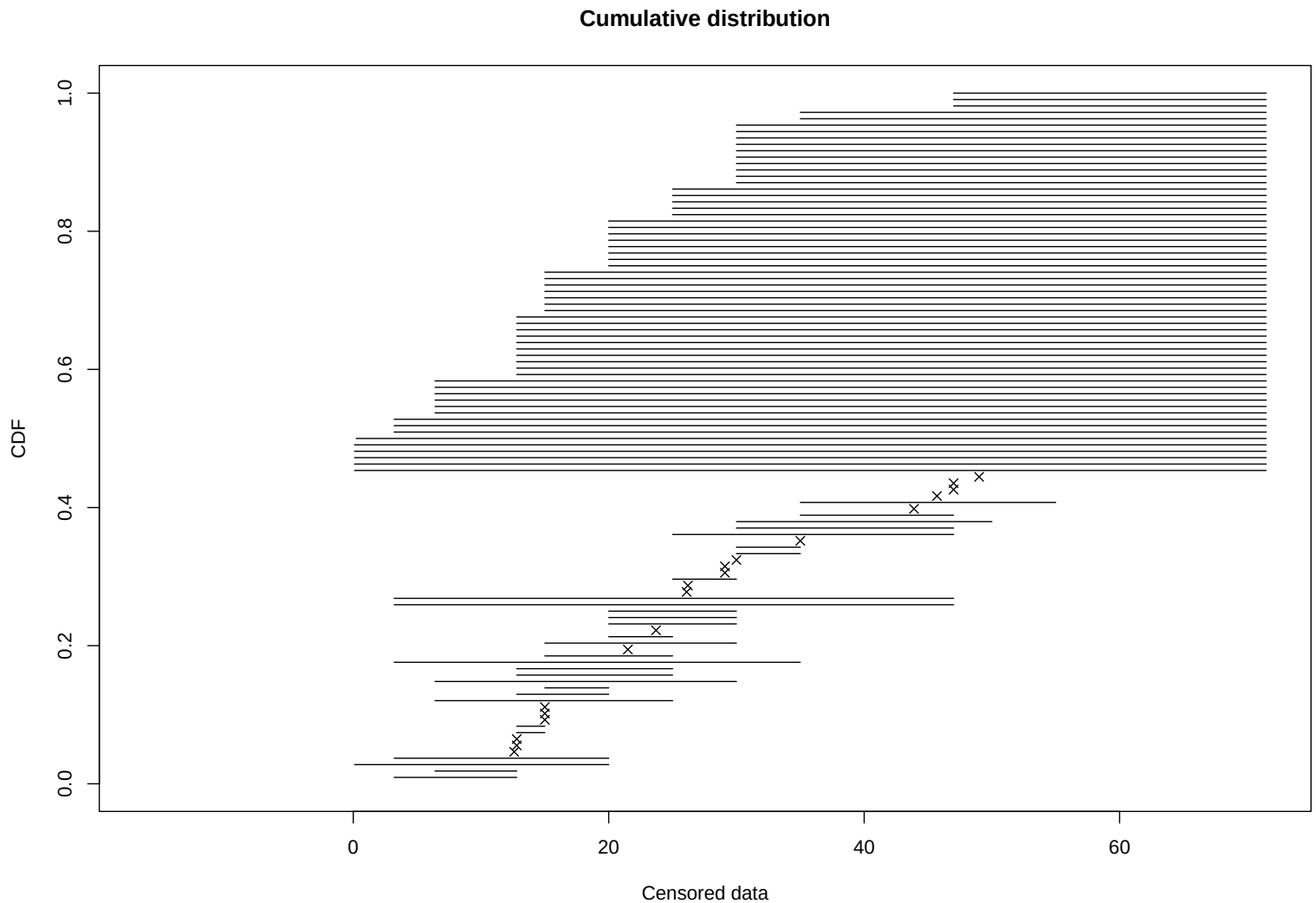
- lewostronne,
- prawostronne,
- przedziałowe.

Dane cenzurowane muszą być odpowiednio przygotowane. Przykładem może być zbiór **salinity**.

```
str(salinity)
```

```
'data.frame': 108 obs. of 2 variables:
$ left : num 20 20 20 20 20 21.5 15 20 23.7 25 ...
$ right: num NA NA NA NA NA 21.5 30 25 23.7 NA ...
```

Dane cenzurowane



Rysunek 5: Wynik wywołania funkcji **plotdistcens** dla zbioru **salinity**.

Dane cenzurowane

Dopasowane odbywa się poprzez estymację parametrów metodą największej wiarogodności dla funkcji wiarogodności dla danych cenzurowanych. Funkcja ta ma postać:

$$L(\theta) = \prod_{i=1}^{N_{nC}} f(x_i|\theta) \cdot \prod_{i=1}^{N_{lC}} F(x_j^u|\theta) \cdot \prod_{i=1}^{N_{rC}} (1-F(x_i^l|\theta)) \cdot \prod_{i=1}^{N_{pC}} (F(x_i^u|\theta)-F(x_i^l|\theta))$$

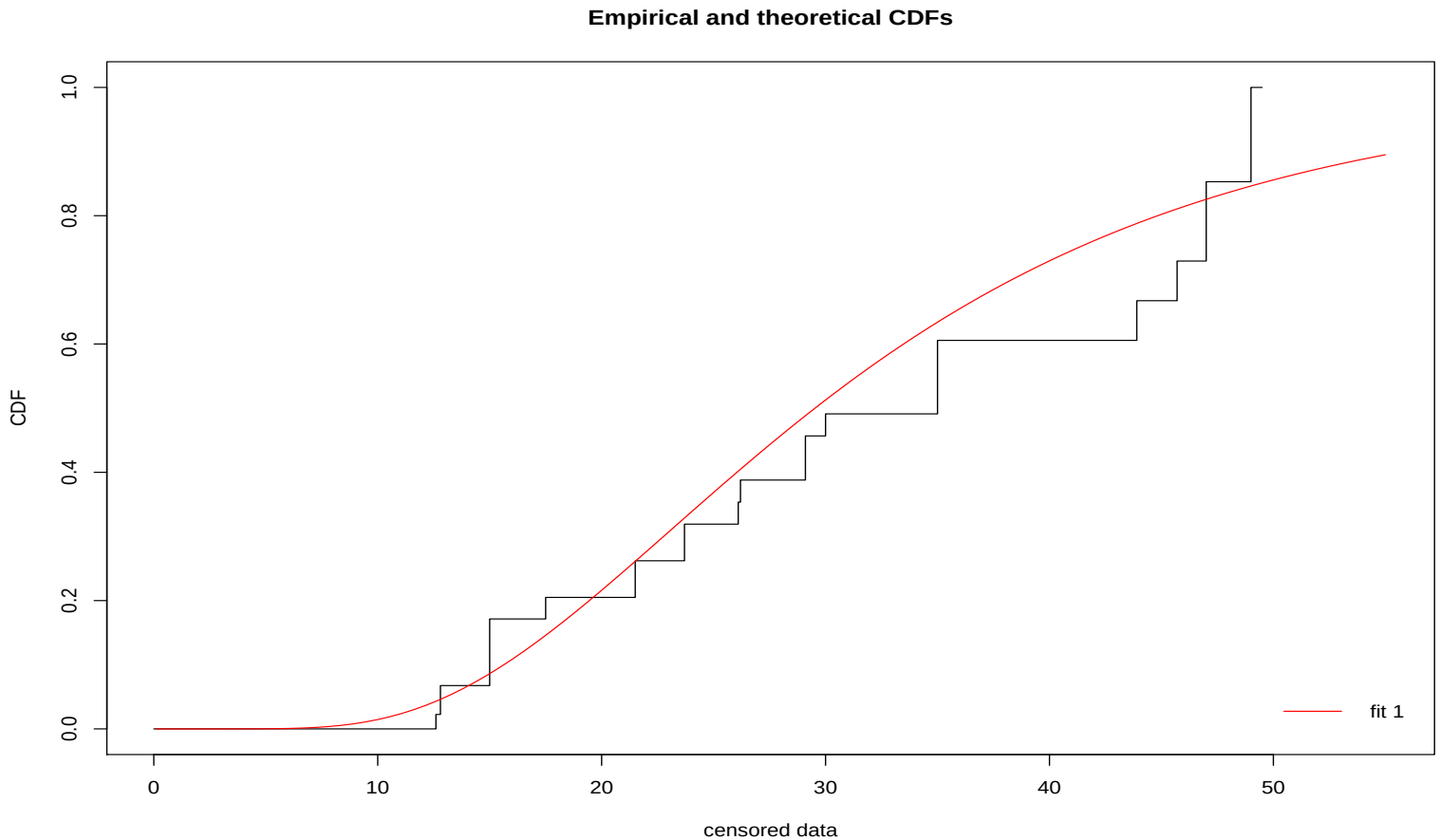
gdzie:

- $N_{nC}, N_{lC}, N_{rC}, N_{pC}$ - odpowiednio liczby obserwacji niecenzurowanych, cenzurowanych lewostronnie, prawostronnie, przedziałowo,
- x_i, x_i^u, x_i^l - i-ta obserwacja, bądź jej górna (u) lub dolna (l) granica,
- f, F - odpowiednio gęstość i dystrybuanta parametrycznego rozkładu.

Dane cenzurowane

```
fsal.ln <- fitdistcens(salinity, "lnorm")
summary(fsal.ln)

Fitting of the distribution ' lnorm ' By maximum likelihood
on censored data
Parameters
        estimate Std. Error
meanlog   3.3854    0.06487
sdlog     0.4961    0.05455
Loglikelihood:  -139.1   AIC:  282.1   BIC:  287.5
Correlation matrix:
        meanlog  sdlog
meanlog  1.0000 0.2938
sdlog    0.2938 1.0000
```

Dane cenzurowane



Rysunek 6: **cdfcompcens** dla danych cenzurowanych i dopasowanego rozkładu rysuje dystrybuanty empiryczną i teoretyczną.

Bibliografia

1. Delignette-Muller M.L., Dutang C., *fitdistrplus: An R Package for Fitting Distributions*, Journal of Statistical Software, 2015(64).

Dziękuję za uwagę!